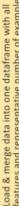
DATA INGESTION Step 1 -

- features and representative number of examples
- Check of shape, indexes, duplicated cols & rows, and if you created new rows/columns with NA's





DATA CLEANING - Step 3 -

Rows/Col with too many NA

Optionally, Remove/or label samples/rows & columns with too many NA, or only NA



Do you see any samples with more then average number of zeros/NAs/anomalies/novelties

Anomaly Detection

DATASET SUMMARY

- Step 2 -



ENCODING Step 4 -

- Step 5 -

ASSUMPTIONS CHECK MODEL

FEATURE ENGINEERING & F. SELECTION Step 6 -

IIP: transform nominal var's into ordinal using the target var. to include them in the analysis

Use: histogram, boxlot, QQ-plot + stat tests

Does it have Normal distribution?

Check Target Variable

should be used to improve residuals disrib.

Check if scaling, or log-transformation

+ skewness, kurtosis, mean, median

Identify incorrect values: Eg: -100 foot size, (plots, histograms, get examples) Next, prepare dict./note/script with (i) feature name, (ii) incorrect values, and

The goal is to have consistent, and non-ambiguous encoding in entire dataset

(target, quantitative, ordinal, nominal,

datetime, spatial, ID columns)

How many variables/samples? What are feature dtypes, and are they

correctly encoded?

Missing Data

What is in my data?

Dataset Overview

1. Define Variable Groups

For quantitative variables

Define Encoding policy

iii) actions to take to either Remove,

Replace or clip identified incorrect

First. Focus on Xi~dependent variable, to

Pair-wise comparisons -

Perform Correlation analysis

identify predictors correlated with the

target variable (use: barplot, table)

Define how many classes, and how

For ordinal variables

If exported from Excel (at any point) check for numerical values turned into

Make sure date-time variables are

properly formatted

Set proper dtype for each feature

2 Cleaning & Organisation

Format dtypes

How much of NA do you have in the dataset - per row/col, and eg: how many rows/col have >=90% of missing

should they be called

integers, means or medians (target

var), and store it in a new column

To help, checking for correlations,

replace class names with sorted

Then, analyse correlations between each

variable pair, to find collinear var's that

heatmap & table with sorted results)

should be removed (use: annotated

- EE
- multivariate normal distribution

use Sklearn functions to create

- transformers for each variable

- imputation strategy variables for one-hot encoding numerical variables that wiii be
- Calculate ViF score for each feature, to detect potential muticolinearity (use: barplot, & table to present the results) one vs rest comparison -

Check Quantitative variables

Define proper encoding for each class

text/object; Tokenize text/object

For nominal Variables

Check if all missing data are encoded

in the same way, if not use one

Identify and remove any duplicates

Remove Duplicates

data-time

What other values can be interpreted missing data ie. which rows/columns

what NA, can be encoded as 0, or

class/category

Feature Values

as NA, eg: 0, or "no value" text

What is the detailed distribution of

(rows & columns)

Missing data

what is the range or class number in

values or replace with classes

Eg: remove inconsistent class names such as "House"/"Home" -> "Home

- Distribution -
- Are they normally distributed? If not, try transformations for each var. eg: log, sqrt, polynomial. Etc... list propositions.
- Extreme Values -
- outliers/zeros in the data? (eg. using z-Can you find outliers or groups of

Encode these variables, eg: age, years/days since last renovation, etc.

properly formatted, (one timezone)

Make sure date-time variables are

For datetime Variables

encoded as NA, and not something else or to remove, rows/columns that

have only missing data

Identify columns with ID, sample order,

Check datetime features

or similar, and check if they are all

The goals, is to have all missing data

Are there any obvious OUTLIERS? How many zero values are there, Check for spacial/datetime variables

Get examples of each feature

class/value distribution,

Which var. have imbalanced

each feature?

No imputation at this stage !

- Do you see any samples with more then average number of zeros, Nas, anomalies scores, on scatter plots)
- Check linear model validity

or novelties

Main question: can we fot linear regression model with our data?

check if residuals/ target values are correlated

Step 3. Check for latent variables

2. Residual analysis Done for selected models with (i) amount of missing data/zeros per sample, (ii) sample order (plot 4) (iii) or the

+ Look for batch effects & autocorrelation,

of extreme values in each sample

Then, use PCA or t-SNE to check if you can

cluster the samples with any spatial var's

Plot target variable against any predictor. Use, scatter plot with smoothed trend-line, or boxplots. Add info on correlation, NA%, mean, median, sd, to each plot

selected, & transformed features should numerical data (int, floats, binary). The GOAL: To generate 2D matrix with only have following characteristics:

- linear relationship with the target No autococelation/batch effects
- Scaled or normalized values No missing data
- Best, No outliers, or identified outliers, + vector with weight

transformers, and add them to one

DEFINE

- Features that you wish to use in the model (see feature selection part in Notebook 2)

- scaled, standardized of normalized, or

CHECK

 How does your transformers treat novelties, especially for nominal variables

REMEMBER TO SAFE

- Feature names, and sample ID for each cell/row/column
 - Dataset & pipeline version

Source data



See Notebook 2



pipeline for testing assumptions and setting up automated functions Here we only establish a simple

Validation Dataset -. used to test the

Feature importance criterion

White test; or Breusch-Pagan test

Print table with models sorted by

Add test a& training error

best performing models

test error, with information on

transformations and features

hyperparameters and data

Test Residuals Homoscedascity Use plot 3 - fannning effect

or Kolmogorov-Smirnov

baseline (RMSE, MAE, R^A2) to select

Bar-plot, or boxplot with errors

calculated for the model/s and

1. Model performance

assessment

Test for residuals autocorelation Use plot 4 – look for patterns

Durbin-Watson Test

tolerate

Results stratification

missing data, novelties, or values outside feature space the model can

Model Resilance; how much of Additional Analyses:

Shapiro-Wilk test, Scipy normaltest, Lilliefors

Mean; Median, Skewness; Kurtosis;

Plot 4. Residuals vs sample order/ID

Test Residuals Normality

Use plot 1 & 2

Plot 3. residuals vs target variable

Plot 1. Histogram
Plot 2. QQ-Plots or PP-Plots

Create Four Standard Plots

ERROR ANALYSIS

- Step 8 -

See Notebook 2

MODEL TRAINING

- Step 7 -

& EVALUTATION